

# OIL PVT CHARACTERISATION USING ENSEMBLE SYSTEMS

MUNIRUDEEN A. OLOSO<sup>1</sup>, MOHAMED G. HASSAN<sup>1</sup>, JAMES BUICK<sup>1</sup>, MOHAMED BADER-EL-DEN<sup>2</sup>

<sup>1</sup>School of Engineering, University of Portsmouth, Portsmouth, United Kingdom

<sup>2</sup>School of Computing, University of Portsmouth, Portsmouth, United Kingdom

E-MAIL: up748008@myport.ac.uk

## Abstract:

In reservoir engineering, there is always a need to estimate crude oil Pressure, Volume and Temperature (PVT) properties for many critical calculations and decisions such as reserve estimate, material balance design and oil recovery strategy, among others. Empirical correlations are often used instead of costly laboratory experiments to estimate these properties. However, these correlations do not always give sufficient accuracy. This paper develops ensemble support vector regression and ensemble regression tree models to predict two important crude oil PVT properties: bubblepoint pressure and oil formation volume factor at bubblepoint. The developed ensemble models are compared with standalone support vector machine (SVM) and regression tree models, and commonly used empirical correlations. The ensemble models give better accuracy when compared to correlations from the literature and more consistent results than the standalone SVM and regression tree models.

## Keywords:

Pressure-Volume-temperature (PVT); empirical, bubblepoint pressure; oil formation volume factor at bubblepoint pressure; ensemble

## 1. Introduction

Prediction of future production or oil reserves from petroleum wells is undoubtedly important for cost-effective operations in the petroleum industry. Production predictions can assist petroleum engineers in economic forecasts; the approach that is often adopted by reservoir engineers is numerical simulation based on log and core analysis results [1]. However, this process can be technically difficult, time consuming and expensive in terms of both labour and computational resources.

An accurate estimation of reservoir fluid pressure, volume, and temperature (PVT) properties, such as bubblepoint pressure, oil formation volume factor, gas solubility, oil viscosity, oil-specific gravity, and gas specific gravity, is essential for many petroleum engineering calculations such as reservoir simulations, recovery estimates, material balance calculations, well completion, facility design

decisions, and production optimization strategies. Practically, the contribution of PVT data is not only limited to the initial stages of the field development but also in future planning and operations. Often, prompt decisions on these issues must be made. Unfortunately, such experimental facilities are not always available. Moreover, conventional PVT laboratory analysis tests can take weeks or even months to complete and have a high associated cost [2].

An alternative to laboratory experimentation is to determine the PVT properties of the crude oil using an empirical method which either uses the Equations of State (EOS), or correlations developed from linear or non-linear regression analysis [1, 2]. The main disadvantage of the EOS is that their inputs include oil compositions which can only be determined by laboratory analyses. On the other hand, correlations usually do not require crude oil compositions as inputs; but at times they fail to capture the uncertainties in the crude oil and do not produce very reliable and consistent predictions.

Several correlations have been developed and the main motivation for seeking new empirical correlation is the inability of the existing ones to properly predict the newly acquired data sets by the authors.

The two crude oil PVT properties that will be examined here are bubblepoint pressure,  $P_b$ , and oil formation volume factor at bubblepoint,  $B_{ob}$ .  $P_b$  is the pressure at which the bubbles of gas first appear and the natural gas contained within the reservoir fluid starts to evolve as a free gas.  $P_b$  is also called saturation pressure because the liquid is saturated with dissolved gas at that stage. Above saturation pressure, the reservoir fluid essentially remains in a liquid phase and all of the volatile components are dissolved in the liquid.

Oil formation volume factor,  $B_o$ , is the volume of reservoir oil required to produce one barrel of oil in the stock tanks. It relates the volume of oil at stock-tank conditions to the volume of oil at elevated pressure and temperature in the reservoir.

This paper has implemented ensemble SVM and regression tree to predict  $P_b$  and  $B_{ob}$ . Both ensemble SVM and ensemble regression tree models have been developed to enable comparison of the two approaches. Single SVM and regression tree models are also implemented and compared with their ensemble models. Average absolute percentage error (AAPRE) and root mean squared error (RMSE) have been used as the criteria in selecting the optimal ensemble models and the base models, unlike most cases where only RMSE is used as the evaluation criterion. The consistency, reliability and accuracy of the results are analysed based on these two error evaluation criteria and the correlation coefficient. The goal of the paper is to demonstrate the applicability of these ensemble methods in predicting PVT properties compared with their single methods especially SVM, which some papers have shown to outperform the common empirical correlations and also some machine learning (ML) algorithms especially neural networks.

The rest of this paper is organised as follows. Section II discusses the empirical and ML methods that have been used to predict PVT properties. General ensemble methods are discussed in section III. The ensemble algorithms developed in this work are highlighted in section IV after a brief discussion of SVM and decision tree. Discussion and comparison of the performances of the developed ensemble models with some selected empirical models are done in section V. Conclusion and direction for future research are presented in section VI.

## 2. Prediction of PVT Properties

A brief literature review of empirical methods and ML techniques that have been used in prediction of PVT properties are explored in this section.

### 2.1. Empirical Correlations for estimating PVT Properties

Many correlations have been developed for the estimation of  $P_b$  and  $B_{ob}$ . Usually, the procedure involves linear and/or non-linear regressions on the available data sets.

Standing [3] developed graphical correlations for both  $P_b$  and  $B_{ob}$  based on 105 sets of experimental data. These graphical correlations were later expressed by equations [4]. The following correlating variables were used for the  $P_b$  correlation: gas solubility ( $R_s$ ), gas gravity ( $\gamma_g$ ), oil API gravity ( $\gamma_{API}$ ) and reservoir temperature (T). For the  $B_{ob}$  correlation, the correlating parameters were:  $R_s$ ,  $\gamma_g$ , oil gravity ( $\gamma_o$ ) and T. On this basis, many other correlations

have been developed for both  $P_b$  and  $B_{ob}$  [2, 5-24].

### 2.2. Machine Learning (ML) for predicting Properties

Generally, the drawback of empirical correlations lie in their high dependency on the range of data used to develop the correlations. To overcome this constraint, some ML techniques have been used for PVT characterisation to improve the prediction of the PVT properties. Artificial Neural Network (ANN) and its variants have been used and have been reported to show improved performance in predicting PVT properties [25-27].

Some notable shortcomings of ANN, such as high training time to reach the optimal result and its 'black box' approach, have resulted in the use of other ML techniques such as Support Vector Machines (SVM) regression, Functional Networks (FN), Genetic Algorithm (GA) and different hybrid systems, among others [28-31].

An ML solution usually involves minimization of the error in the learning algorithm. It is notable that many of the machine learning algorithms work by performing a local search that may become stuck in a local minima which causes the model to perform poorly when presented with new data [32]. Also, there is always more than one possible final model that can be achieved from training a chosen ML algorithm based on a given data set. Hence, there is a need to combine decisions from different predictors. This is achieved in ensemble ML algorithms.

Ensemble ML based on SVM and decision tree algorithms is demonstrated in this paper for prediction of PVT properties. These models are evaluated using both AAPRE and RMSE. As there is no consensus on which of these two error evaluating criteria is the best [33], the two criteria have been combined. Each model is assigned two different ranks based on AAPRE and RMSE. The ranks are assigned in ascending order of both AAPRE and RMSE. The base model or the ensemble with the least value in either case will have the highest rank. As a rule of thumb, a consistent and stable model is expected to have the same ranks for both AAPRE and RMSE.

## 3. Ensemble Machine Learning

Ensemble ML is a combination of multiple base models-classifiers or regressors. Each base model covers a different part of the input space or the complete input space. Though there is no definitive taxonomy for building the ensemble models, some successful approaches and

methodologies have been widely adopted [32]. Popular among these methods are bagging and boosting. Another notable method, mainly for classification problem is ADABOOST [34].

In ensemble ML, each base model is usually trained on a slightly different training set and the predictions from all the models are combined with the goal of producing a better and more accurate output than the individual base models. The ensemble models aim to reduce the expected errors in the predicted target output values. Expected error consists of the prediction bias and the variance [35]. In other words, it is the expected difference between the estimated function and the true function.

The prediction error from high bias and low variance in some base models (causing under fitting of the training data sets) or low bias and high variance in other base models (causing over fitting of the training data sets) can be reduced when different base models are combined.

Bagging [36], which is also known as bootstrap aggregation involves training multiple models with training sets of data randomly drawn with replacement from the base training data sets. The training data sets for the base models are called *bootstraps*. Hence, bagging involves training different models with different samples and usually predictions are obtained by averaging the results of the different base models for a regression problem.

Boosting involves training and improving a weak learning algorithm into a strong one [37]. In boosting, the training dataset for each subsequent model increasingly focuses on instances wrongly predicted by the previous weaker model. ADABOOST (adaptive boosting algorithm) is one of the most used boosting algorithms which automatically adapts to the data given to it.

#### 4. Proposed Ensemble Models

Two different ensemble models are implemented in this work. The first ensemble models uses decision tree as the learning method while the second one uses SVM regression.

SVM is a statistical machine learning method that generates input-output mapping functions from a set of training data. It uses the principle of structural risk minimization, seeking to minimise the upper bound of the generalisation error rather than just minimising the training error. In a simple pattern recognition problem, SVM uses a linear separating hyperplane to create a classifier with a

maximal margin [38]. When the input cannot be linearly transformed (e.g. complex classification problem or regression problem), SVM first nonlinearly transforms the input space into a higher-dimensional feature space. The transformation is achieved by using nonlinear mapping functions which are generally referred to as kernel functions. Typical kernel functions include RBF, Gaussian and polynomial functions.

A decision tree is of two types: classification and regression trees. Classification trees apply to problems where the output data is discrete while regression trees apply to problems where the output is continuous. Recursive partitioning (RP) is usually used for growing regression trees. The computational complexity of the RP used for growing regression trees is highly dependent on the choice of the best split for a given node. Algorithms proposed for implementation of classification and regression trees include Automatic Interaction Detection (AID), THeta Automatic Interaction Detection (THAID), Classification and Regression Trees (CART), C4.5, CHi-squared Automatic Interaction Detector (CHAID), Fast and Accurate Classification Tree (FACT). Detailed review of these methods have been done by Wei-Tin Loh [39].

##### 4.1 Ranking Selection Criteria

As earlier mentioned, two important error measurements are used as evaluation criteria in this paper (AAPRE and RMSE).

The ensemble regression model and the base models for the ensemble. No priority is given to either error measurement as there is no consensus on which one is superior [34]. Rather, a rule of thumb is adopted which is exemplified in the Table 1.

TABLE 1: Example of the Ranking Selection

Model Number	AAPRE Rank	RMSE Rank	Decision
Model_1	1	1	Accepted
Model_2	3	2	Rejected
Model_3	2	3	Rejected
Model_4	4	4	Accepted
Model_5	5	5	Accepted

#### 4.2 Ensemble Regression Trees

A regression tree can be considered to be a variant of decision tree which maps the input space into a real valued domain. In regression tree, the target output consists of numeric or continuous values (real numbers) [40]. In this work, an ensemble regression tree was created using least squared Boosting (LSBoost) and is shown as Algorithm 1.

The aim of this algorithm is to search within a large space of different inputs to dynamically determine the optimal parameters for both the regression tree and the ensemble average.  $X$  is the input data set and  $Y$  is the output vector. Both  $X$  and  $Y$  must have the same number of rows.  $n$  is the maximum number of branch nodes for each decision tree while  $m$  is the desired maximum number of trees in an ensemble model.

1. Select randomly  $x$  data sets as 70% of the entire data sets ( $X$ ) and the corresponding  $y$  from the output ( $Y$ )
2. Do for  $k=1$  to  $n$ , number of possible splits (e.g.  $n=100$ )
3. Do for  $j=1$  to  $m$ , number of trees in each ensemble (e.g.  $m=1000$ )
4. Initialize a regression tree template,  $T_k$
5. Create ensemble  $F(j, k)$  using LSBoost method
6. Compute RMSE and AAPRE of the predicted target using  $F(j, k)$
7. Continue until  $j=m$
8. Continue until  $k=n$

Algorithm 1. Procedure for the Ensemble Regression Trees

The implementation begins by randomly selecting  $x$  data set from the entire  $X$  data.  $K$  is set as the number of splits or branches in each regression tree.  $j$  determines the number of regression trees in each ensemble model.  $T_k$  creates a regression tree with maximum  $k$  splits.  $F(j, k)$  creates an ensemble regression tree using the LSBoost function in MATLAB. Each ensemble model is evaluated using AAPRE and RMSE. The ensemble model with minimal AAPRE and RMSE is adopted as the final model. All the generated ensemble models are ranked separately based on AAPRE and RMSE in ascending orders. Any model whose ranks for both AAPRE and RMSE are not the same will be eliminated. The adopted ensemble model is the one that has equal highest ranks for both AAPRE and RMSE.

#### 4.3. Ensemble Support Vector Machine Regression

SVM is a statistical learning theory originally developed by Cortes and Vapnik for classification problems [41]. SVM uses the principle of structural risk minimization, seeking to minimize an upper bound of the generalization error.

For a linear problem, it uses a linear separating hyperplane to create a classifier with a maximal margin. For a regression problem, or where the input classes cannot be linearly separated in the original input space, the SVM first nonlinearly transforms the original input space into a higher dimensional space where a maximal separating hyperplane is constructed. Kernel function is used to create the required SVM hyperplane. Possible kernels includes Polynomial, linear, RBF and Gaussian. This is shown as Algorithm 2.

In algorithm 2,  $X$  and  $Y$  are the inputs and the output respectively.  $C$  is the penalty factor which should neither be too large nor too small to prevent over-fitting and under-fitting respectively [42]. However, there is no definite criteria for determining its value as it is problem dependent. The value of  $\varepsilon$  controls width of the  $\varepsilon$ -insensitive zone of the SVM regression model and determines its accuracy.

1. Select randomly  $x$  data sets as 70% of the entire data sets ( $X$ ) and the corresponding  $y$  from the output ( $Y$ )
2. Iterate for  $C=1$  to  $N$
3. Iterate for kernel,  $k \rightarrow \{RBF, Gaussian\}$
4. Iterate for  $\varepsilon \in \{10^{-4}, 10^{-5}, 10^{-6}\}$
5. Compute each SVM model  $F(C, k, \varepsilon)$
6. Evaluate each SVM model using AAPRE and RMSE
7. Continue for the next  $\varepsilon$
8. Continue for the next  $k$
9. Continue until  $C=N$
10. Choose the best  $n$  models for the ensemble
11. Predict the testing target  $Y$  from the testing input  $X$  using the  $n$  base SVM models
12. Compute the ensemble output  $\frac{1}{n} \sum_{i=1}^n \hat{Y}_i$ , where  $\hat{Y}_i$  is the predicted target by the  $i$ th SVM base model

Algorithm 2: Ensemble Support Vector Machine Regression

The three main parameters that control each SVM model are  $C$ ,  $k$  and  $\varepsilon$ . Each SVM model  $F(C, k, \varepsilon)$  is evaluated using AAPRE and RMSE. To create the final ensemble SVM model,  $n$  number of the optimized base

models will be selected as the base models. Before selecting the  $n$  base SVM models, ranks are assigned to all the base SVM models based on both AAPRE and RMSE in ascending orders. Similarly, any model whose ranks for both AAPRE and RMSE are not the same will be eliminated. The output of the ensemble SVM will be the average of the outputs of the optimized SVM base models.

## 5. Ensemble Learning of PVT Properties

### 5.1. Data Sets

A total of 895 data points were available for the simulation. 327 data points were collected from different published papers [8, 10, 11, 43], while the remaining are unpublished PVT data. Table 2 shows the statistical representation of the data sets. These data points cover PVT reports mainly from the Gulf of Mexico, Malaysia, the Middle East and Africa.

### 5.2. Simulations

Algorithm 1 was implemented to develop the ensemble regression tree models for both  $P_b$  and  $B_{ob}$  while algorithm 2 was implemented to develop the ensemble SVM regression.  $N$  was chosen as 1000 while  $n$  was chosen as 5 in both instances of  $P_b$  and  $B_{ob}$  modelling. All simulations were done using MATLAB. For both  $P_b$  and  $B_{ob}$  predictions, the inputs were: gas/oil ratio, temperature, stock-tank oil gravity ( $API$ ) and gas gravity.

The results of the ensemble models for both investigated PVT properties,  $P_b$  and  $B_{ob}$ , have been compared with the commonly used empirical correlations [3, 5, 8] and some recently developed ones [2, 24]. Two commonly used performance measures for PVT predictions have also been adopted. These are correlation coefficient ( $R^2$ ), AAPRE and RMSE. The correlation coefficient lies between 0 and 1. A correlation coefficient of 1 indicates a perfect correlation and 100% accuracy of the values predicted by the model. On the other hand, lower AAPRE and RMSE indicates better performance and strong accuracy of the model in predicting the target values. Mathematically, correlation coefficient ( $R^2$ ), AAPRE and RMSE are given as follows.

$$R^2 = 1 - \frac{[\sum_{i=1}^n (Y_{pred} - Y_{exp})_i]^2}{\sum_{i=1}^n [(Y_{exp} - \bar{Y})_i]^2} \quad (1)$$

Where

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n (Y_{exp})_i \quad (2)$$

$$AAPRE = \frac{1}{n} \sum_{i=1}^n |E_i| \quad (3)$$

where

$$E_i = \left( \frac{Y_{exp} - Y_{pred}}{Y_{exp}} \right) \times 100 \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_{pred} - Y_{exp})^2} \quad (5)$$

$Y_{exp}$  is the experimental value of the data point and  $X_{pred}$  is the corresponding predicted value.

Table 2. Statistical Description of the PVT data

PVT Property	Minimum value	Maximum value
Bubblepoint Pressure (psia)	122.5	6523
Oil formation volume factor at bubblepoint (bbl/STB)	1.032	2.493
Gas/Oil ratio (scf/STB)	26	2315
Temperature (F)	74	288
Stock-tank oil gravity (API)	16.3	53.2
Gas gravity (air=1)	0.5781	1.367

### 5.3. Results and Discussion

For bubblepoint prediction, Table 3 shows the results for the ensemble SVM regression and ensemble regression tree models along with some recognised correlations [3, 5, 8], recently developed correlations [2, 24], standalone SVM and regression tree models. Though the standalone SVM model has the highest correlation coefficient,  $CC$ , but it is inconsistent in the error evaluation as its AAPRE (12.512%) is higher than that of the ensemble SVM regression model.

The ensemble SVM regression model has the second highest  $CC$  (0.9839; approximately equal to that of standalone SVM model) and lowest AAPRE (12.437%) and RMSE (194.150) values. The ensemble SVM model shows more reliable and consistent results than the standalone SVM model. The second best consistent results are from the ensemble regression tree model, while the poorest performance is given by the Jarahin *et al's* [2] correlation which has the lowest  $CC$  and highest AAPRE and RMSE

values.

Table 3. Statistical Measures for Bubblepoint Pressure Prediction (Testing Data)

Model	CC	AAPRE	RMSE
Standing [3]	0.9129	16.430	449.574
AlMarhoun [8]	0.9553	13.930	411.645
Vazquez and Beggs [5]	0.9215	18.842	520.370
Arabloo <i>et al</i> [24]	0.9171	15.259	435.312
Jarahin <i>et al</i> [2]	0.8285	75.029	1750.900
Regression Tree	0.9474	15.039	346.341
SVM	0.9840	12.512	193.649
Ensemble Regression Tree (this work)	0.9815	13.126	195.581
Ensemble SVM Regression (this work)	0.9839	12.437	194.150

In Table 4, the results for the developed ensemble models along with some published correlation equations and standalone SVM and regression tree models for  $B_{ob}$  prediction are shown. In this case the Ensemble Regression Tree model has the best and consistent accuracy with highest CC of 0.9951 and lowest prediction errors based on AAPRE and RMSE with values of 0.9487 and 0.0212% respectively.

Table 4. Statistical measures for Prediction of Oil FVF at bubblepoint pressure (Testing Data)

Model	CC	AAPRE	RMSE
Standing [3]	0.9832	2.0555	0.0434
AlMarhoun [8]	0.9881	1.7826	0.0362
Vazquez and Beggs [5]	0.9728	2.7656	0.0557
Arabloo <i>et al</i> [24]	0.9881	1.7391	0.0345
Regression Tree	0.9771	2.2978	0.0476
SVM	0.9941	1.2273	0.0235
Ensemble Regression Tree (this work)	0.9951	0.9487	0.0212
Ensemble SVM Regression (this work)	0.9948	1.1842	0.0226

Notably, the standalone regression tree model has a poorer accuracy compared to its ensemble model. Also, the ensemble SVM model has better accuracy than the standalone SVM model. The least accuracy is given by the

correlation of Vazquez and Beggs [5], having the least CC (0.9728) and highest AAPRE (2.7656%) and RMSE (0.0557).

#### 4. Conclusions

Ensemble SVM regression and ensemble regression tree models have been developed in this paper for the prediction of two important crude oil PVT properties: bubblepoint pressure and oil formation volume factor at bubblepoint pressure. For bubblepoint pressure prediction, the ensemble SVM regression model gave the most reliable performance while the ensemble regression tree model gave the highest accuracy in predicting the oil formation volume factor at the bubblepoint pressure.

Clearly, the two developed ensemble models in this work outperformed all the common empirical correlations and the recently developed correlations which were considered for comparison. This shows that the developed ensemble models have better capabilities in capturing the uncertainties embedded in the crude oil data. Hence, there is a strong potential for the ensemble systems developed here to replace the empirical correlations that are still widely used in the petroleum industry for estimating these two PVT properties.

The consistency and reliability of ensemble models over their respective individual models have also been demonstrated. This also stresses one of the aims and advantages of ensemble models where the base models complement one another with the potential of reducing bias and prediction error.

In the future, the developed ensemble models can be applied to other PVT properties such as viscosity, gas/oil ratio and isothermal oil compressibility.

#### Acknowledgements

Oloso Munirudeen thanks the Petroleum Technology Development Fund, Nigeria, for sponsoring his PhD research at the University of Portsmouth. The authors also thank Professor T.A Blasingame of Texas A& M University for sharing some PVT data with them.

## References

- [1] Oloso, M.A., Khoukhi, A., Abdulraheem, A., and Elshafei, M.: 'Prediction of crude oil viscosity and gas/oil ratio curves using advances to neural networks'. Proc. SPE/EAGE Reservoir Characterization and Simulation Conference, Abu Dhabi, UAE, 19-21 October, 2009.
- [2] Jarrahan, A., Moghadasi, J., and Heidaryan, E.: 'Empirical estimating of black oils bubblepoint (saturation) pressure', *Journal of Petroleum Science and Engineering*, 2015, 126, pp. 69-77.
- [3] Standing, M.B.: 'A pressure-volume-temperature correlation for mixtures of California oils and gases', in *Drilling and Production Practice*, 1947.
- [4] Standing, M.B.: 'Volumetric and phase behavior of oil field hydrocarbon systems' (Dallas: Society of Petroleum Engineers, 1977).
- [5] Vazquez, M., and Beggs, H.D.: 'Correlations for fluid physical property prediction', *Journal of Petroleum Technology*, 1980, 32, (06), pp. 968-970.
- [6] Lasater, J.: 'Bubble point pressure correlation', *Journal of Petroleum Technology*, 1958, 10, (05), pp. 65-67.
- [7] Glasø, Ø.: 'Generalized Pressure-Volume-Temperature Correlations'. JPT 32 (5) 785-795," SPE-8016-PA. DOI: 10.2118/8016-PA1980.
- [8] Al-Marhoun, M.A.: 'PVT correlations for Middle East crude oils', *Journal of Petroleum Technology*, 1988, 40, (05), pp. 650-666.
- [9] Kartoatmodjo, T.R., and Schmidt, Z.: 'New correlations for crude oil physical properties', Society of Petroleum Engineers, 1991.
- [10] Dokla, M., and Osman, M.: 'Correlation of PVT Properties for UAE Crudes (includes associated papers 26135 and 26316)', SPE Formation Evaluation, 1992, 7, (01), pp. 41-46.
- [11] Omar, M., and Todd, A.: 'Development of new modified black oil correlations for Malaysian crudes', Proc. SPE Asia Pacific Oil and Gas Conference, Singapore, 8-10 February, 1993.
- [12] Velarde, J., Blasingame, T., and McCain Jr, W.: 'Correlation of black oil properties at pressures below bubble point pressure-A new approach', In *Annual Technical Meeting. Petroleum Society of Canada*, Calgary, Alberta, June 8 - 11, 1997.
- [13] Almehaideb, R.: 'Improved PVT correlations for UAE crude oils', Proc. Middle East Oil Show and Conference, Bahrain, 15-18 March, 1997.
- [14] Hanafy, H., Macary, S., ElNady, Y., Bayomi, A., and El Batanony, M.: 'A new approach for predicting the crude oil properties', Proc. SPE Production Operations Symposium, Oklahoma City, Oklahoma, 9-11 March, 1997.
- [15] Petrosky Jr, G., and Farshad, F.: 'Pressure-volume-temperature correlations for Gulf of Mexico crude oils', SPE Reservoir Evaluation & Engineering, 1998, 1, (05), pp. 416-420.
- [16] Khairy, M., El-Tayeb, S., and Hamdallah, M.: 'PVT correlations developed for Egyptian crudes', *Oil and Gas Journal*, 1998, 96, (18).
- [17] Al-Shammasi, A.: 'A review of bubblepoint pressure and oil formation volume factor correlations', SPE Reservoir Evaluation & Engineering, 2001, 4, (02), pp. 146-160.
- [18] Elsharkawy, A.M.: 'An empirical model for estimating the saturation pressures of crude oils', *Journal of Petroleum Science and Engineering*, 2003, 38, (1), pp. 57-77.
- [19] Valko, P., and McCain, W.: 'Reservoir oil bubblepoint pressures revisited; solution gas-oil ratios and surface gas specific gravities', *Journal of Petroleum Science and Engineering*, 2003, 37, (3), pp. 153-169.
- [20] Dindoruk, B., and Christman, P.G.: 'PVT Properties and Viscosity Correlations for Gulf of Mexico Oils', SPE Reservoir Evaluation & Engineering, 2004, 7, (06), pp. 427-437.
- [21] Malallah, A., Gharbi, R., and Algharaib, M.: 'Accurate estimation of the world crude oil PVT properties using graphical alternating conditional expectation', *Energy & fuels*, 2006, 20, (2), pp. 688-698.
- [22] Hemmati, M.N., and Kharrat, R.: 'A correlation approach for prediction of crude oil PVT properties', Proc. SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain, 11-14 March, 2007.
- [23] Khamsehchi, E., Rashidi, F., Rasouli, H., and Ebrahimian, A.: 'Novel empirical correlations for estimation of bubble point pressure, saturated viscosity and gas solubility of crude oils', *Petroleum Science*, 2009, 6, (1), pp. 86-90.
- [24] Arabloo, M., Amooie, M.-A., Hemmati-Sarapardeh, A., Ghazanfari, M.-H., and Mohammadi, A.H.: 'Application of constrained multi-variable search methods for prediction of PVT properties of crude oil systems', *Fluid Phase Equilibria*, 2014, 363, pp. 121-130.

- [25] Gharbi, R., and Elsharkawy, A.M.: 'Neural network model for estimating the PVT properties of Middle East crude oils'. SPE Reservoir Evaluation & Engineering, vol. 2, pp. 255 - 265, June, 1999.
- [26] Al-Marhoun, M., and Osman, E.: 'Using artificial neural networks to develop new PVT correlations for Saudi crude oils'. Proc. Abu Dhabi International Petroleum Exhibition and Conference, Abu Dhabi, United Arab Emirates, 13-16 October, 2002.
- [27] Osman, E.-S.A., and Al-Marhoun, M.A.: 'Artificial neural networks models for predicting PVT properties of oil field brines'. Proc. SPE Middle East Oil and Gas Show and Conference, Kingdom of Bahrain, 12-15 March, 2005.
- [28] El-Sebakhy, E.A., Sheltami, T., Al-Bokhitan, S.Y., Shaaban, Y., Raharja, P.D., and Khaeruzzaman, Y.: 'Support vector machines framework for predicting the PVT properties of crude oil systems'. Proc. SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain, 11-14 March, 2007.
- [29] Khoukhi, A.: 'Hybrid soft computing systems for reservoir PVT properties prediction', Computers & Geosciences, 2012, 44, pp. 109-119.
- [30] Talebi, R., Ghiasi, M.M., Talebi, H., Mohammadyan, M., Zendehboudi, S., Arabloo, M., and Bahadori, A.: 'Application of soft computing approaches for modeling saturation pressure of reservoir oils', Journal of Natural Gas Science and Engineering, 2014, 20, pp. 8-15.
- [31] Asoodeh, M., and Bagheripour, P.: 'Estimation of bubble point pressure from PVT data using a power-law committee with intelligent systems', Journal of Petroleum Science and Engineering, 2012, 90, pp. 1-11.
- [32] Dietterich, T.G.: 'Ensemble methods in machine learning': 'Multiple classifier systems' (Springer, 2000), pp. 1-15.
- [33] Chai, T., and Draxler, R.R.: 'Root mean square error (RMSE) or mean absolute error (MAE)?-Arguments against avoiding RMSE in the literature', Geoscientific Model Development, 2014, 7, (3), pp. 1247-1250.
- [34] Freund, Y., and Schapire, R.E.: 'Experiments with a new boosting algorithm'. In ICML, vol. 96, pp. 148-156. 1996.
- [35] Friedman, J., Hastie, T., and Tibshirani, R.: 'The elements of statistical learning' (Springer series in statistics Springer, Berlin, 2001. 2001).
- [36] Breiman, L.: 'Bagging predictors', Machine learning, 1996, 24, (2), pp. 123-140.
- [37] Schapire, R.E.: 'The strength of weak learnability', Machine learning, 1990, 5, (2), pp. 197-227.
- [38] Kecman, V.: 'Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models' (Bradford Book, 2001. 2001)
- [39] Loh, W.Y.: 'Fifty years of classification and regression trees', International Statistical Review, 2014, 82, (3), pp. 329-348.
- [40] Lior Rokach, O.M.: 'Data Mining With Decision Trees: Theory and Applications (2nd Edition)' (World Scientific Publishing Company, 2015. 2015).
- [41] Cortes, C., and Vapnik, V.: 'Support-vector networks', Machine learning, 1995, 20, (3), pp. 273-297.
- [42] Alpaydin, E.: 'Introduction to machine learning' (MIT press, 2014. 2014).
- [43] Bello, O., Reinicke, K., and Patil, P.: 'Comparison of the performance of empirical models used for the prediction of the PVT properties of crude oils of the niger delta', Petroleum Science and Technology, 2008, 26, (5), pp. 593-609.